Aufgabenstellung 3. Teilleistung Gruppenarbeit / GitHub

Bitte bearbeitet diese Aufgabe in einer **Gruppe von 4-6 Personen**. Eure Gruppenmitglieder findet ihr auf Moodle.

Bitte nutzt für die Bearbeitung der im folgenden geschilderten Aufgabe GitHub und Git und die in zugehörigen Sitzung erlernten Inhalte. Die Bewertung dieser Teilleistung wird an der Anwendung der erlernten Inhalte (sowohl vorgestellte Funktionen als auch der Übungsaufgabe) festgemacht. Eure Skripte sollen **ohne Fehler durchlaufen und die Aufgabestellung richtig umsetzen**, allerdings wird darüber hinaus das Programmieren nicht streng bewertet (die Dokumentation der Funktionen und Skripte allerdings schon!). Wichtig für die Bewertung ist, dass ihr als **ganze Gruppe** unter Beteiligung **aller Mitglieder** die Aufgabe über ein **GitHub Repository** bearbeiten und auf diesem erkennbar ist, dass ihr über einen **längeren Zeitraum gemeinsam** an der Aufgabe gearbeitet habt und damit die **in vorgestellten Workflows** zur Verwendung von GitHub eingesetzt habt.

Abzugeben ist über Moodle für eure Gruppe **ein Link zu eurem GitHub Repository**. Stellt sicher, dass ich dieses einsehen kann. Anhand von diesem wird die Bewertung stattfinden. **Abgabefrist für den Link: 25.02.2024, 23.59h**.

In Moodle ist wird ein Datensatz „titanic.csv“ zur Verfügung gestellt. Dieser enthält die nachfolgenden Informationen über die Passagiere auf der Titanic:

- PassengerID: ID-Variable
- Survived: Hat den Untergang der Titanic überlebt? Ja (1), Nein (0)
- Pclass: Klasse des Reisenden (ordinal mit 1 > 2 > 3)
- Name: Name des Reisenden
- Sex: Geschlecht (male/female)
- Age: Alter in Jahren beim Untergang (Für Kleinkinder auch in Dezimalzahlen)
- SibSp: Anzahl an Geschwistern und Ehefrauen an Bord
- Parch: Anzahl an Eltern und Kinder an Bord
- Ticket: Ticketnummer
- Fare: Ticketpreis
- Cabin: Kabinennummer
- Embarked: Zustiegshafen (C = Cherbourg; Q = Queenstown; S = Southampton)

Über euer GitHub Repository unter Verwendung von Git und in gleichbeteiligter Gruppenarbeit, bearbeitet bitte die folgenden Aufgaben:

1. **Zwei** Gruppenmitglieder erstellen ein R-Skript, mit welchem der Titanic-Datensatz eingelesen und für eine Analyse vorverarbeitet wird. Am Ende des Skriptes sollen nur noch sinnvolle Variablen im Datensatz vorhanden sein, die auch bei einer Analyse sinnvoll verwendet werden können. Bearbeitet hierfür die folgenden Schritte:
    - Extrahiert aus dem Namen eine Variable mit der Anrede der Person, d.h. „Mr.“, „Mrs.“, „Mse.“ usw., damit später fehlende Werte im Alter ersetzt werden können. Beachtet hierbei, dass gewisse Anreden wie „Ms.“, „Miss.“ oder „Mlle“ inhaltlich gleichbedeutend sind (in diesem Beispiel eine junge, unverheiratete Frau). Die Anrede „Master“ bezeichnet einen kleinen Jungen.

- Codiert die Variablen „Survived“, „Sex“, „Embarked“ als factor um.
- Überführt die Variable „Pclass“ in einen ordered-factor.
- Imputiert fehlende Werte in der Variable „Age“ mithilfe der erzeugten Variable „Anrede“ über ein Imputationsverfahren eurer Wahl (z.B. arithmetisches Mittel, Median, usw.)
- Extrahiert aus der Variable „Cabin“ die folgenden Informationen und erzeugt neue Variablen hierfür:
  - Backbord oder Steuerbord? Tipp: Kabinen mit einer ungeraden Nummer liegen auf Steuerbord, die anderen auf Backbord.
  - Deck: Vorangehender Buchstabe der Kabinennummer
  - Einträge mit unbekannter Kabinennummer, d.h. „“ setzt ihr auf NA.
- Entfernt am Ende die Variablen „PassengerID“, „Name“, „Ticket“ und „Cabin“ aus dem Datensatz
- Speichert das R-Skript, sowie den neuen Datensatz in dem GitHub-Repository ab.

2. **Gemeinsam als ganze Gruppe** erstellt zwei weitere R-Skripte. Im 4. Schritt sollen diejenigen Gruppenmitglieder, die nicht an (1.) gearbeitet haben, den Datensatz analysieren (Deskription und Visualisierung). Hierzu sollen nun Funktionen erstellt werden, die dabei genutzt werden.
   a. Funktionen-R-Skript 1 soll (mindestens) folgende Funktionen enthalten:
      i. Eine Funktion, die verschiedene geeignete deskriptive Statistiken für metrische Variablen berechnet und ausgibt
      ii. Eine Funktion, die verschiedene geeignete deskriptive Statistiken für kategoriale Variablen berechnet und ausgibt
      iii. Eine Funktion, die geeignete deskriptive bivariate Statistiken für den Zusammenhang zwischen zwei kategorialen Variablen berechnet ausgibt
      iv. Eine Funktion, die geeignete deskriptive bivariate Statistiken für den Zusammengang zwischen einer metrischen und einer dichotomen Variablen berechnet und ausgibt
      v. Eine Funktion, die eine geeignete Visualisierung von drei oder vier kategorialen Variablen erstellt
      vi. Freiwillig: weitere zur Deskription und Visualisierung geeignete Funktionen
   b. Funktionen-R-Skript 2 soll Helfer-Funktionen enthalten, die nicht selbst zur Deskription und Visualisierung der Daten verwendet werden, sondern die nur in Funktionen-Skript 1 Anwendung finden (→ interne Funktionen). Funktionen-R-Skript 2 muss mindestens eine Funktion enthalten.
3. Denkt auch an eine gute Dokumentation aller Funktionen. Nutzt euer GitHub Repository um darüber zu diskutieren, welche Funktionen sinnvoll und notwendig sind.
4. **Diejenigen Gruppenmitglieder, die nicht an (1.) gearbeitet haben**, sollen mit Hilfe der in (3.) in Funktionen-R-Skript 1 erstellten Funktionen den aufgeräumten Datensatz aus (1.) analysieren (Deskription und Visualisierung). Hierzu soll ein weiteres Skript im Repository erstellt werden. Hierbei sollte das R-Skript im Minimum jede der Funktionen (i) bis (vi) aus Funktionen-R-Skript 1 einmal anwenden. Denkt bei der Aufgabe über sinnvolle Analysen nach, z.B. Überlebensrate gegen andere Variablen. Wie verhält sich der Ticketpreis?

5. Diskutiert anschließend im GitHub Repository als ganze Gruppe die Ergebnisse. Möglicherweise haben die Gruppenmitglieder, die an (1.) gearbeitet haben noch weitere Ideen.